Published in final edited form as:
J Am Stat Assoc. 2018 ; 113(523): 1243–1254. doi:10.1080/01621459.2017.1330204.

Quantile-Optimal Treatment Regimes

Lan Wang [Professor],
School of Statistics, University of Minnesota, Minneapolis, MN 55455

Yu Zhou [Graduate student],
School of Statistics, University of Minnesota, Minneapolis, MN 55455

Rui Song [Associate Professor], and
Department of Statistics, North Carolina State University, Raleigh, NC 27695

Ben Sherwood [Assistant Professor]
School of Business, University of Kansas

Abstract

Finding the optimal treatment regime (or a series of sequential treatment regimes) based on individual characteristics has important applications in areas such as precision medicine, government policies and active labor market interventions. In the current literature, the optimal treatment regime is usually defined as the one that maximizes the average benefit in the potential population. This paper studies a general framework for estimating the quantile-optimal treatment regime, which is of importance in many real-world applications. Given a collection of treatment regimes, we consider robust estimation of the quantile-optimal treatment regime, which does not require the analyst to specify an outcome regression model. We propose an alternative formulation of the estimator as a solution of an optimization problem with an estimated nuisance parameter. This novel representation allows us to investigate the asymptotic theory of the estimated optimal treatment regime using empirical process techniques. We derive theory involving a nonstandard convergence rate and a non-normal limiting distribution. The same nonstandard convergence rate would also occur if the mean optimality criterion is applied, but this has not been studied. Thus, our results fill an important theoretical gap for a general class of policy search methods in the literature. The paper investigates both static and dynamic treatment regimes. In addition, doubly robust estimation and alternative optimality criterion such as that based on Gini's mean difference or weighted quantiles are investigated. Numerical simulations demonstrate the performance of the proposed estimator. A data example from a trial in HIV+ patients is used to illustrate the application.

Keywords

dynamic treatment regime; nonstandard asymptotics; optimal treatment regime; precision medicine; quantile criterion

1 Introduction

A treatment regime can be described as a function from the space of covariates to the set of treatment options. Depending on the application, a treatment can represent a drug, a device, a program, a policy, an intervention or a strategy. The problem of estimating an optimal treatment regime has recently received considerable attention. Medical doctors have long been interested in tailoring a patient's medical treatment according to the individual's unique genetic information, health history, environmental exposure, needs and preferences. Economists are interested in finding the most effective active labor market programs (job search training, computer training, etc.) for an unemployed job seeker (Frölich (2008), Behncke et al. (2009), Staghøj et al. (2010), Wunsch (2013)). In political science, researchers are interested in selecting the best strategies (personal visits, phone calls, mailings, etc.) to increase voter turnout (Gerber and Green (2000), Imai and Ratkovic (2013)).

Existing work on estimating an optimal treatment regime has mainly focused on the mean-optimal treatment regime, which if followed by the whole population would yield the largest average outcome (assuming a larger outcome is preferable). Popular approaches for estimating mean-optimal treatment regimes include model-based methods such as Q-learning (Watkins and Dayan, 1992; Murphy, 2005b; Chakraborty et al., 2010; Moodie and Richardson, 2010; Goldberg and Kosorok, 2012; Song et al., 2015), A-learning (Robins et al., 2000; Murphy, 2003, 2005a), and model-free or policy search methods (Robins and Rotnitzky, 2008; Orellana and Robins, 2010; Zhang et al., 2012a; Zhao et al., 2012, 2015a). Other relevant work includes Robins (2004); Moodie et al. (2007, 2009); Henderson et al. (2010); Cai et al. (2011); Qian and Murphy (2011); Thall et al. (2011); Imai and Ratkovic (2013); Huang et al. (2015); Tao and Wang (2017), among others. We refer to the recent books (Chakraborty and Moodie, 2013; Kosorok and Moodie, 2016) and review articles (Qian et al., 2012; Chakraborty and Murphy, 2014; Laber et al., 2014; Schulte et al., 2014; Wallace and Moodie, 2014) for a more comprehensive list of references. In econometrics, an independent line of interesting work explored a decision theory framework for estimating statistical treatment rules (Manski, 2004; Dehejia, 2005; Hirano and Porter, 2009; Stoye, 2009; Bhattacharya, 2009; Bhattacharya and Dupas, 2012; Tetenov, 2012).

In a variety of applications, criteria other than the mean (or the average) may be more sensible. When the outcome has a skewed distribution (e.g., survival time of patients), it may be desirable to consider the treatment regime that maximizes the median of the distribution of the potential outcome. Sometimes, the tail of the potential outcome distribution is of direct importance. When evaluating government job training programs to improve earnings, policy makers may ask which program does best to improve earnings on the lower tail. An optimal treatment regime with respect to the tail criterion is even more attractive if the sacrifice is little at the central part of the potential outcome distribution as compared to the mean-optimal treatment regime. A simple numerical example illustrating phenomenon of this nature is given in Section 2. The same numerical example also reveals that the mean-optimal treatment regime may work poorly (or even have detrimental effect) at the tails.

In this paper, we study a general framework for estimating the quantile-optimal treatment regime in both static and dynamic settings, the latter of which involves estimating a sequence of treatment regimes that may vary over time based on a longitudinal study. Given a class of treatment regimes, we consider a robust estimator of the quantile-optimal treatment regime that does not require specifying an outcome regression model. By now, it has been widely recognized (Qian and Murphy, 2011; Zhang et al., 2012a; Zhao et al., 2012; Matsouaka et al., 2014; Zhao et al., 2015b) that a fundamental challenge in estimating the optimal treatment regime is specifying a reliable outcome model, which describes how the treatment and covariates influence the outcome and how they interact with each other. A misspecified outcome model can result in biased estimation of the optimal treatment regime. The difficulty of specifying outcome models is more pronounced when estimating the optimal dynamic treatment regime using longitudinal data, for which model-based approaches would require specifying a sequence of outcome models, one for each decision point. However, complete nonparametric estimation of optimal treatment regimes suffers from the curse of dimensionality and does not provide easy-to-interpret treatment regimes.

Although some recent work has made important contributions to estimating the optimal treatment regime without an outcome model (Robins and Rotnitzky, 2008; Robins et al., 2000; van der Laan et al., 2005; Orellana and Robins, 2010; Zhang et al., 2012a, 2013; Zhao et al., 2012, 2015b), they have considered only the mean-optimal criterion and have not studied the asymptotic distribution of the estimated optimal treatment regime. In fact, as will be shown later in the paper, the classical asymptotic theory does not apply to this class of estimators even for the mean-optimal criterion.

We propose a novel formulation of the estimator as a solution of an optimization problem with an estimated nuisance parameter. This representation allows us to further investigate the asymptotic theory of the estimated optimal treatment regime using empirical processes techniques. Our study reveals that the theory involves nonstandard asymptotics. We have rigorously established that: (1) the estimated parameter indexing the quantile-optimal treatment regime converges at a cube-root rate to a nonnormal limiting distribution that is characterized by the maximizer of a centered Gaussian process with a parabolic drift; and (2) the value function corresponding to the quantile optimal treatment regime can be estimated at an $O_p(n^{-1/2})$ rate. This new framework is broad in the sense that it also provides an alternative formulation of the mean optimal criterion, for which the same type of nonstandard asymptotics would arise. Thus, we fill an important gap in the literature. Moreover, the framework can be adapted to alternative criteria such as those based on weighted quantile or Gini's mean difference (Section 1.2 of online supplement). The main practical advantage of the proposed estimator is that it circumvents the difficulty of specifying a reliable outcome regression model, which has undue influence on estimating the optimal treatment regime. We also investigate doubly robust estimation (Section 1.1 of online supplement), which can incorporate an outcome regression model when it is available.

In the causal inference context, several authors have considered estimating the quantile treatment effects for comparing several pre-determined treatment regimes (Rubin, 1974; Rosenbaum and Rubin, 1983; Hogan and Lee, 2004; Chernozhukov and Hansen, 2005; Zhang et al., 2012b). These authors have not investigated the fundamental problem of estimating the optimal treatment regimes in the quantile framework, which is much more complex than estimating the quantile specific treatment effect when the treatment assignment is given. Potentially, the recent work on discrete Q-learning in Moodie et al. (2014) can be applied to first estimate the probabilities and then invert them to estimate quantiles, but this application has not been systematically studied. Linn et al. (2015) independently considered estimating quantile-optimal treatment regime. However, their approach depends on applying threshold interactive model-based Q-learning at a sequence of thresholding values and then performing inversion. The method requires specifying the underlying outcome models and is computationally intensive even for homoscedastic error outcome models. Furthermore, Linn et al. (2015) has not studied the asymptotic theory we considered here.

The rest of the paper is organized as follows. The quantile-optimal treatment regime is proposed in Section 2. The estimation procedure and asymptotic distribution are introduced in Section 3. Section 4 investigates quantile-optimal dynamic treatment regimes. Simulation studies and a data example are reported in Section 5. Section 6 considers doubly robust estimation and alternative optimality criteria. The proofs are given in the Appendix. Additional technical details and numerical results can be found in the online supplement. The methods proposed in this paper can be implemented using the R package *quantoptr* (Zhou et al., 2017).

2 Quantile-optimal treatment regime

Let A be the binary variable denoting treatment (0 or 1 corresponding to two treatment options), and let Y denote the outcome. Without loss of generality, we assume that a larger value of the outcome is preferable. To evaluate the treatment effect, we consider the potential or counterfactual outcome framework (Neyman (1990), Rubin (1978)) for causal models. Let $Y^*(1)$ be the potential outcome had the subject been assigned to treatment 1; and $Y^*(0)$ be the potential outcome had the subject been assigned to treatment 0. For each individual in the sample, we observe either $Y^*(1)$ or $Y^*(0)$, but not both. It is assumed that the observed outcome is $Y = Y^*(1)A + Y^*(0)(1 - A)$, that is, the observed outcome is the potential outcome corresponding to the treatment the subject actually receives. This is often referred to as the consistency assumption in causal inference. We also adopt the stable unit treatment value assumption (Rubin (1986)), that is, a subject's outcome of receiving a treatment is not influenced by the treatments received by other subjects.

Let X denote an l-dimensional vector of covariates. A treatment regime is defined as a function $d(X)$, that maps the covariates vector X to the set of treatment options, here $\{0,1\}$. For example, $d(X) = I(X \le 3/5)$ would assign a subject with $X = 0.2$ to treatment 1. Given treatment regime $d(X)$, the corresponding potential outcome is $Y^*(d) = Y^*(1)d(X) + Y^*(0)(1 - d(X))$, that is, $Y^*(d)$ is the outcome one would observe if a subject with covariate value X is assigned to treatment 1 or 0 following treatment regime $d(X)$. We assume that $(Y^*(1), Y^*(0))$ is independent of A conditional on X (unconfoundedness assumption, Rosenbaum and Rubin (1983)), which is automatically satisfied in randomized trials.

Given a collection $\mathbb{D}$ of treatment regimes, the optimal treatment regime is typically defined as the one that maximizes the average of the potential outcome: $E(Y^*(d))$. Here, we consider a new quantile-optimal treatment regime, which is defined as

$$\arg\ max_{d \in \mathbb{D}} Q_{\tau}(Y^{*}(d)), \quad (1)$$

where $\tau \in (0, 1)$ is the quantile level of interest and $Q_{\tau}\ (Y^{*}(d))$ is the τth quantile of $Y^{*}(d)$, specifically, $Q_{\tau}(Y^{*}(d)) = \inf\ \{t : F^{*}(t) \geq \tau\}$ with F^{*} denoting the distribution function of $Y^{*}(d)$.

To illustrate how the quantile-optimal treatment regime differs from the mean-optimal treatment regime, we consider a simple but instructive example. The outcome, Y_i, satisfies $Y_i = 1 + 3A_i + X_i - 5A_iX_i + (1 + A_i + 2A_iX_i)\varepsilon_i$, where $\varepsilon_i \sim N(0,1)$ $X_i \sim$ Uniform[0, 1], and $A_i = 1$ (or 0) if subject i receives treatment (or control). We consider the following six treatment regimes: (1) $A_i = 0,\ \forall\ i$; (2) $A_i = I(X_i \leq 3/5)$; (3) $A_i = I(X_i < 1/2)$; (4) $A_i = I(X_i < 1/5)$; (5) $A_i = I(X_i \leq 1/10)$; (6) $A_i = 1,\ \forall\ i$; and (7) random assignment $P(A_i = 1) = 0.5$. It is easy to derive that treatment regime 2 is the mean-optimal treatment regime. Table 1 summarizes the mean, the 0.25 quantile ($Q_{0.25}$) and 0.10 quantile ($Q_{0.10}$) of the potential outcome distribution corresponding to each of the six treatment regimes, based on a Monte Carlo experiment with 10^6 observations. We observe that regime 3 is the best if one is interested in maximizing the first quartile of the potential outcome distribution; whereas regime 4 performs best with respect to the 0.10 quantile. If we consider the hypothetical setting where the outcome is the survival time of cancer patients, then regime 2 (mean-optimal treatment regime) may have detrimental effect for patients at the lower tail, corresponding to weaker patients. Regime 3 significantly improves the survival time of the patients at the lower tail, while its mean value is comparable to that of regime 2. Thus, regime 3 is preferable if doctors wish to improve the life span of more severely ill patients without sacrificing the average treatment benefit of the population.

3 Estimation and large sample theory

3.1 Estimating quantile-optimal treatment regime

To explain the idea, we first consider a randomized trial with two treatment options (denoted by 1 and 0). Extensions to observational studies and dynamic treatment regimes will be discussed later. The observed data $\{X_i,\ Y_i,\ A_i\}$, $i = 1, \ldots, n$, are independent and identically distributed copies of $\{X,\ Y,\ A\}$. Our aim is to estimate the quantile-optimal treatment regime given a class of feasible treatment regimes $\mathbb{D} = \left\{I(X^{T}\beta > 0) : \beta \in \mathbb{B}\right\}$ where β indexes different treatment regimes and $\mathbb{B}$ is a compact subset of $\mathbb{R}^{l}$. This class of *single-index* decision rules has been popular in practice (Zhang et al., 2012a, 2013; Zhao et al., 2012) due to its simplicity and interpretability. It is straightforward to show that this class contains the mean-optimal treatment regime corresponding to some popular choices of outcome models. For example, for the outcome model $E(Y|A,\ X) = \beta_0 + \beta_1X_1 + \beta_2X_2 + A(\beta_3 + \beta_4X_1 + \beta_5X_2)$, the corresponding mean-optimal treatment regime is $I(\beta_3 + \beta_4X_1 + \beta_5X_2 > 0)$. An alternative class of treatment regimes that are practically appealing is the class of *thresholding rules* of the form $I(X_1 > \beta_1, \ldots, X_1 > \beta_1)$, for some constants $\beta_1, \ldots, \beta_1$. Even for these relatively simple forms, asymptotic theory for the estimated optimal treatment regime, no matter what the criterion is, is nontrivial. It is worth pointing out that it is not necessary that the class of candidate treatment regimes includes the theoretically global optimal treatment regimes, as the interpretability of the treatment regime is often of fundamental importance.

We will focus on the single-index treatment regimes, as the theory for the thresholding decision rules is similar and simpler. Given a β, $\in \mathbb{B}$, let $d(X, \beta) = I\{X^T\beta > 0\}$ be the treatment regime indexed by β, which is sometimes denoted by d_β for notational simplicity. For a quantile level of interest τ $(0 < \tau < 1)$, we would like to estimate β_0 = arg $max_{\beta \in B} Q_\tau(Y^*(d\beta))$, the parameter indexing the quantile-optimal treatment regime. To do so, we make use of an induced missing data framework motivated by Zhang et al. (2012a). Let $C(\beta) = Ad(X, \beta) + (1 - A)(1 - d(X, \beta))$. In the induced missing data framework, the "full data" of interest, but not completely observed, are $\{Y^*(d_\beta), X\}$; and the observed data are $\{C(\beta), C(\beta)Y^*(d_\beta), X\} = \{C(\beta), C(\beta)Y, X\}$. If $C(\beta) = 1$, then potential outcome $Y^*(d_\beta)$ is observed; if $C(\beta) = 0$ then Y^* $(d\beta)$ is "missing". Furthermore, $Y^*(d_\beta)$ and $C(\beta)$ are independent conditional on X. Thus, the induced missing data structure satisfies the missing at random assumption. Let

$$\hat{Q}_\tau(\beta) = \underset{a}{\operatorname{argmin}}\, n^{-1} \sum_{i=1}^{n} C_i(\beta)\rho_\tau(Y_i - a), \quad (2)$$

where $\rho_\tau(u) = u\,(\tau - I(u < 0))$ is the quantile loss function. As stated in the following lemma (proof given in the online supplement), $\hat{Q}_\tau(\beta)$ is a consistent estimator of the τth quantile of $Y^*(d_\beta)$.

Lemma 1—If condition ($C1$) in Section 3.3 is satisfied, then we have $\hat{Q}_\tau(\beta) \to Q_\tau(Y^*(d_\beta))$ in probability, $\forall\ \beta \in \mathbb{B}$.

The estimator for β_0 that corresponds to the quantile-optimal treatment regime is

$$\hat{\beta}_n = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}}\, \hat{Q}_\tau(\beta). \quad (3)$$

The estimated quantile-optimal treatment regime is $d_{\hat{\beta}} = I(X^T\hat{\beta} > 0)$. Section 2.1 of the online supplement provides the calculation details.

3.2 Alternative formulation of the proposed estimator

As the treatment regimes involve indicator functions, the nonsmoothness leads to nonstandard asymptotics even when the mean criterion is used. The asymptotic theory is challenging and involves a cube-root convergence rate and a non-normal limiting distribution, see Section 3.3 for details. Even for the mean criterion, the asymptotic distribution theory of the estimated optimal treatment regime has not yet been systematically developed in the literature.

To facilitate the development of theory, we introduce a novel reformulation that represents the quantile-optimal treatment regime parameter estimator (3) as a solution of an optimization problem with an estimated nuisance parameter. To motivate the reformulation, let

$$g(\cdot, \beta, m) = C(\beta)I\{Y - m > 0\}, \quad (4)$$

$$m_0 = \sup\left\{m: \sup_{\beta \in \mathbb{B}} Pg(\cdot, \beta, m) \geq (1 - \tau)/2\right\}, \quad (5)$$

$$\beta_0 = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}}\ Pg(\cdot, \beta, m_0). \quad (6)$$

The function $g(\cdot, \beta, m)$ is motivated by the first-order condition of the maximization problem in (2). For a randomized trial, $P(C(\beta) = 1 \mid X) = P(C(\beta) = 0 \mid X) = \frac{1}{2}$. Thus, $P(g(\cdot, \beta, m)) = \frac{1}{2}P(Y^*(d_\beta) > m)$, which is equal to $\frac{1 - \tau}{2}$ if $m = Q_\tau(Y^*(d_\beta))$. For any given β, because $g(\cdot, \beta, m)$ is monotonically decreasing in m, it follows that $Q_\tau(Y^*(d_\beta))$ is the largest value of m such that $Pg(\cdot, \beta, m)$ is greater than or equal to $\frac{1 - \tau}{2}$. Therefore, m_0 defined in (5) is the largest achievable τth quantile of $Y^*(d_\beta)$ over $\beta \in \mathbb{B}$; and β_0 defined in (6) is the population value of the parameter that indexes the optimal treatment regime.

Now, let P_n denote the empirical expectation, that is, $P_n f(Z) = n^{-1}\sum_{i=1}^n f(Z_i)$, where $Z_1, \ldots, Z_n$ is a random sample and $f(\cdot)$ is an arbitrary function. Then, $\widehat{m}_n = \sup\left\{m: \sup_{\beta \in \mathbb{B}} P_n g(\cdot, \beta, m) \geq (1 - \tau)/2\right\}$ is the estimator of the largest achievable τth quantile of $Y^*(d_\beta)$ over the class of treatment regimes under consideration. We have the following alternative expression of the estimator in (3):

$$\widehat{\beta}_n = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}}\ P_n g(\cdot, \beta, \widehat{m}_n). \quad (7)$$

In other words, $\widehat{\beta}_n$ is the value of β at which the supremum of $P_n g(\cdot, \beta, \widehat{m}_n)$ is achieved, thus it is the estimator of the parameter that indexes the optimal treatment regime. This reformulation was partly motivated by the least median of squares estimator of Rousseeuw (1984). A benefit of this reformulation is that we also obtain the convergence rate of $\widehat{m}_n$, which is the estimator for the maximally achievable value function (here, the maximally achievable τth quantile of the potential outcome) as a by product (see Lemma 2 in Section 3.3).

3.3 Asymptotic properties

We assume the following regularity conditions.

(C1) Potential outcomes $Y^*(1)$ and $Y^*(0)$ both have continuous distributions with bounded, continuously differentiable density functions.

(C2) The population parameter indexing the optimal treatment regime, $\beta_0 \in \mathbb{R}^l$, which satisfies $\| \beta_0 \| = 1$, where $\| \cdot \|$ denotes the Euclidean norm, is unique and is an interior point of $\mathbb{B}$, a compact subset of the parameter space.

(C3) X has a continuously differentiable density function $f(\cdot)$. The angular components of X, considered as a random element of the unit sphere $\mathbb{S}$ in $\mathbb{R}^l$ has a bounded, continuous density with respect to the surface measure on $\mathbb{S}$.

(C4) Let $q(X, \delta) = S_{1,X}(m_0 + \delta) - S_{0,X}(m_0 + \delta)$, where $S_{1,X}(\cdot)$ and $S_{0,X}(\cdot)$ denote the conditional survival functions of $Y^*(1)$ and $Y^*(0)$ give X, respectively; and $\dot{q}(X, 0)$ and $\dot{f}(X)$ denote the gradients with respect to X. The $l \times l$ matrix $V = \frac{1}{2} \int I\left\{X^T \beta_0 = 0\right\}(f(X)\dot{q}(X,0) + q(X,0)\dot{f}(X))' \beta_0 X X^T d\sigma$ is positive definite, where σ is the surface measure on the hyperplane $\{X : X^T \beta_0 = 0\}$.

Condition (C1) is a standard assumption on the potential outcomes in causal inference. Condition (C2) is an identifiabilitv condition for β_0. Conditions (C3) and (C4) are technical conditions to evaluate the quadratic drift and covariance function of the Gaussian process that are used to characterize the asymptotic distribution of $\hat{\beta}_n$. The matrix V in (C4) characterizes the quadratic drift of the Gaussian process. These two conditions are similar to those in Example 6.4 of Kim and Pollard (1990). In particular, condition (C3) is mainly imposed for the convenience of calculating the derivative of the surface integral in the proof of Lemma 2. It can be relaxed to allow some of the components of X to be discrete at the expense of a more complex expression for V. The new formulation in Section 3.2 connects the problem of estimating β_0 to the class of estimation problems with cube root asymptotics (Kim and Pollard, 1990). However, the result of Kim and Pollard (1990) is not directly applicable because our estimator of β_0 contains an estimated nuisance parameter $\widehat{m}_n$. Lemma 2 below shows that $\hat{\beta}_n$ nearly maximizes the objective function in (7) in which $\widehat{m}_n$ is replaced by the limiting value m_0.

Lemma 2—Under conditions (Cl)–(Cf),

(1) $\widehat{m}_n = m_0 + O_p(n^{-1/2})$.

(2) $P_n g(\cdot, \hat{\beta}_n, m_0) \geq \sup_{\beta \in \mathbb{B}} P_n g(\cdot, \beta, m_0) - o_p(n^{-2/3})$.

The first part of Lemma 2 shows that $\widehat{m}_n$ has a root-n convergence rate. This result is of independent interest as it tells us how well we could estimate the theoretically largest achievable value of the criterion function. The details of the derivation of the lemma are given in the Appendix. Lemma 2 confirms that $\hat{\beta}_n$ nearly maximizes $P_n g(\cdot, \beta, m_0)$. This allows us to further derive the asymptotic distribution of $\hat{\beta}_n$, which is expressible as a functional of two-sided Brownian motion with a quadratic drift. This result is stated in the following theorem.

Theorem 1—Assume conditions (*C1*)–(*C4*) are satisfied. Then, $n^{1/3}(\hat{\beta}_n - \beta_0)$ converges in distribution to $\arg\max_t Z(t)$, where the process $Z(t) = -\frac{1}{2}t^T Vt + W(t)$, V is an $l \times l$ positive definite matrix and $W(t)$ is a centered Gaussian process with continuous sample paths and covariance kernel function $K(\cdot, \cdot)$. The expressions for V and $K(\cdot, \cdot)$ are given in the proof of the theorem in the Appendix.

Remark 1: If the mean-optimal criterion is of interest, then we let $g^*(\cdot, \beta, \mu) = C(\beta)(Y - \mu)$ and $\hat{\mu}_n = \sup\left\{\mu : \sup_{\beta \in \mathbb{B}} P_n g^*(\cdot, \beta, \mu) > 0\right\}$. The estimated parameter indexing the mean-optimal treatment regime has the representation $\hat{\beta}_n^{\text{mean}} = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}}\, P_n g * (\cdot, \beta, \hat{\mu}_n)$. It is straightforward to adapt the techniques developed in this paper to show that the estimated parameter indexing the mean-optimal treatment regime has a non-standard convergence rate and a non-normal limiting distribution. This fills an important gap in the literature.

Remark 2: If the observed data arise from observational studies, the above formulation and theory can be extended using propensity score weighting. For observational studies, we have $Y^* (d_\beta) \perp C(\beta)|X$, which is guaranteed under the common causal inference assumption $\{Y^*(1), Y^*(0)\} \perp A|X$. Thus, the "missing at random" assumption is satisfied in the induced missing data framework of Section 3.1. Let $\pi(X) = P(A = 1|X)$, then the propensity score $P(C_\beta = l|X)$ has the expression $\pi(X)d(X, \beta) + (1 - \pi(X))(1 - d(X, \beta))$. We denote the propensity score by $\pi_c(X, \beta)$ for notational simplicity. We then estimate the τth quantile of $Y^*(d_\beta)$ by $\tilde{Q}_\tau(\beta) = \underset{a}{\operatorname{argmin}}\, n^{-1}\sum_{i=1}^{n} \frac{C_i(\beta)}{\hat{\pi}_c(X_i, \beta)}\rho_\tau(Y_i - a)$, where $\hat{\pi}_c(X_i, \beta)$ is an estimator of the propensity score $\pi_c(X, \beta)$. A simple way to obtain $\hat{\pi}_c(X_c, \beta)$ is to estimate $\pi(X)$ based on $\{A_i, X_i\}$, $i = 1, \ldots, n$, using logistic regression, which models $\pi(X)$ as $\pi(X, \gamma) = \exp(X^T\gamma)/(1 + \exp(X^T\gamma))$. One may also use semiparametric or nonparametric models, which renders greater flexibility but demands heavier computation. The estimated parameter indexing the quantile-optimal treatment regime is given by $\underset{\beta \in \mathbb{B}}{\operatorname{argmax}}\, \tilde{Q}_\tau(\beta)$.

4 Quantile-optimal dynamic treatment regimes

When treating chronic medical conditions, it is frequently necessary to vary the treatment (e.g., drug type, dose) over time according to how the patient responds to the previous treatment. This motivates us to consider estimating the quantile-optimal dynamic treatment regime (DTR) using data from longitudinal studies, which can also help catch possible delayed treatment effects. Comparing with the static treatment regime discussed earlier, a new challenge is the presence of time-dependent covariates that may be simultaneously confounders and intermediate variables.

Consider a two-stage longitudinal study in which the subject receives treatment $A_1 \in \{0,1\}$ at stage 1 and treatment $A_2 \in \{0, 1\}$ at stage 2. We are interested in the outcome at the end of the study. We would like to estimate the optimal DTR $d = (d_1, d_2)$, where d_j can depend on the realized covariates and treatment history before the jth decision, $j = 1, 2$. The baseline vector of covariates is denoted by X_1, the potential outcomes are denoted by $\left\{X_2^*(d_1), Y^*(d)\right\}$, where $X_2^*(d_1)$ is the covariate information between decisions d_1 and d_2 had the subject received treatment d_1; and $Y^*(d)$ is the potential outcome had the subject received treatment $d = (d_1, d_2)$. As before, we define the quantile-optimal DTR as $d^{\text{opt}} = \underset{d \in \mathbb{D}}{\operatorname{argmax}}\ Q_\tau(Y^*(d))$. $H_1 = \{X_1\}$ and $H_2 = \{X_1, A_1, X_2\}$. We adopt the no unmeasured confounder or sequential ignorability assumption (Robins (1997)), that is, given any regime (a_1, a_2), $A_1 \perp \left\{X_2^*\left(a_1\right), Y^*\left(a_1, a_2\right)\right\} \mid H_1$ and $A_2 \perp Y^*(a_1, a_2) | H_2$. In other words, treatment A_j received in the jth stage ($j = 1, 2$) is independent of any future (potential) covariate or outcome conditional on the history. We also adopt the positivity assumption, that is, there exist positive constants $c_1 < c_2$ such that $c_1 \le P(A_j = a|H_j) < c_2$, with probability one, for $a \in \{0, 1\}$ $j = 1.2$. Assume that the class of candidate treatment regimes is indexed by $\xi = (\beta^T, \gamma^T)^T \in \mathbb{B} = \mathbb{B}_1 \times \mathbb{B}_2, d_\xi = (d_\beta, d_\gamma)$, where $d_\beta(H_1) = I(H_1^T\beta > 0)$and $d_\gamma(H_2) = I(H_2^T\gamma > 0)$.

The observed data are denoted by $\{X_{i1}, A_{i1}, X_{i2}, A_{i2}, Y_i\}$, $i = 1, \ldots, n$, where X_{i1} denotes the baseline vector of covariates for subject i, A_{i1} is the treatment subject i receives at stage 1, X_{i2} denotes the vector of intermediate information observed between the two stages, A_{i2} is the treatment subject i receives at stage 2, and Y_i is the observed outcome for subject i (as before, a larger value is preferred). To estimate the optimal treatment regime, we consider a similar induced missing data structure, as motivated by Zhang et al. (2013). For a given treatment regime d_ξ, the "full data" are $(X_1, X_2^*(d_\beta), Y^*(d_\xi))$. Let $C_\xi = \infty$ if $A_1 = d_\beta$ and $A_2 = d_\gamma$. In this case, $(X_1, X_2, Y) = (X_1, X_2^*(d_\beta), Y^*(d_\xi))$, and we observe the potential outcome. Let $C_\beta = 2$ if $A_1 = d_\beta$ but $A_2 \ne d_\xi$ (dropout before decision 2); and let $C_\beta = 1$ if $A_1 \ne d_\beta$ and $A_2 \ne d_\xi$ (dropout before decision 1). Note that this induced missing data structure mimics the monotone dropout scenario for longitudinal data. We can verify that the setup satisfies the missing at random assumption, that is, missingness may be related to the observed information but is conditionally independent of what is missing.

Let $\pi_1(H_1) = P(A_1 = 1 \mid H_1)$ and $\pi_2(H_2) = \pi_2(\bar{X}_2, a_2) = P(A_2 = 1 | \bar{X}_2, a_2)$, where $\bar{X}_2 = (X_1^T, X_2^T)^T$ is an l-dimensional vector. It is important to note that H_2 depends on the treatment received at the first stage. If the subject receives $A_1 = a_1 \in \{0, 1\}$ at the first stage, we sometimes write H_2 as $H_2(a_1) = \{X_1, a_1, X_2\}$ to emphasize the dependence, for which case $X_2 = X_2^*(a_1)$ by the consistency assumption. Similarly, for $A_1 = d_\beta(H_1)$, we sometimes write H_2 as $H_2(d_\beta) = \{d_\beta(X_1), X_2\}$. The potential outcomes correspond to d_ξ are denoted by $\left\{X_1, X_2^*(d_\beta(X_1)), Y^*(d_\xi)\right\}$ or simply $\left\{X_2^*(d_\beta), Y^*(d_\xi)\right\}$.

As before, we would minimize $P_n\left(\frac{I(C_\xi = \infty)}{P(C_\xi \infty | H_2)}\rho_\tau(Y - a)\right)$ to estimate the τth quantile of $Y^*(d_\xi)$. Note that $C_\xi = \infty$ if and only if $A_1 = d_\beta(X_1)$ and $A_2 = \mathrm{d}_\gamma(H_2(d_\beta))$, in other words, $H_2 = H_2(d_\beta)$ or the observed history is the potential history corresponding to d_β. Thus, in the above inverse probability weighted quantile loss function

$$P(C_\xi = \infty | H_2) = P(C_\xi = \infty | X_1, X_2^*(d_\beta(X_1))) = P(A_1 = d_\beta | X_1, X_2^*(d_\beta(X_1)))P(A_1 = d_\gamma | A_1 = d_\beta(X_1), X_1, X_2^*(d_\beta(X_1))) = P(A_1 = d_\beta | X_1), P(A_2 = d_\gamma | H_2(d_\beta))$$

where $P(A_1 = d_\beta | X_1) = [\pi_1(H_1)d_\beta + (1 - \pi_1(H_1))(1 - d_\beta)]$ and $P(A_2 = d_\gamma | H_2(d_\beta)) = [\pi_2(H_2(d_\beta))d_\gamma + (1 - \pi_2(H_2(d_\beta)))(1 - d_\gamma)]$. For notational simplicity, we denote $P(C_\xi = \infty | H_2)$ by $\pi(\xi)$. Formally, the estimate of the τth quantile of $Y^*(d_\xi)$ is given by $\widehat{Q}_\tau(\xi) = \underset{a}{\operatorname{argmin}}\, n^{-1}\sum_{i=1}^n \frac{I(C_{\xi,i} = \infty)}{\pi(\xi)}\rho_\tau(Y_i - a)$, where $C_{\xi,I}$ is the value of C_ξ for subject i. The consistency of $\widehat{Q}_\tau(\xi)$ is established in the online supplement. The estimator of the parameter indexing the optimal DTR from the class $\mathbb{D}$ is defined as $\widehat{\xi} = \underset{\xi = (\beta^T, \gamma^T)^T \in \mathbb{B}}{\operatorname{argmax}} \widehat{Q}_\tau(\xi)$. The estimated quantile-optimal treatment regime is $d_{\widehat{\xi}} = (d_{\widehat{\beta}}, d_{\widehat{\gamma}})$.

In the following, we assume that the data arise from a SMART (sequential, multiple, assignment randomized trials), which has been recommended as a standard design for optimal DTR estimation (Lavori and Dawson, 2000; Murphy, 2008). For a SMART, $\pi_1(H_1)$ and $\pi_1(H_2)$ are both known by design, thus $\pi(\xi)$ is known for any given ξ. Let $g(\cdot, \xi, m) = \frac{I(C_\xi = \infty)}{\pi(\xi)}I(Y > m)$ and $\widehat{m}_n = \sup\left\{m : \sup_\xi P_n g(\cdot, \xi, m) \geq (1 - \tau)\right\}$. We have the following alternative expression $\widehat{\xi}_n = \underset{\xi}{\operatorname{argmax}}\, P_n g(\cdot, \xi, \widehat{m}_n)$. Let $m_0 = \sup\{m : \sup_\xi Pg(\cdot, \xi, m) \geq (1 - \tau)\}$ and $\xi_0 = \underset{\xi}{\operatorname{argmax}}\, Pg(\cdot, \xi, m_0)$. Under similar conditions as for Theorem 1, it can be derived that the limiting distribution of $n^{1/3}(\widehat{\xi}_n - \xi_0)$ is that of the maximizer of a centered Gaussian process with a quadratic drift.

Theorem 2

Under conditions $(C1^*)$–$(C4^*)$ given in the online supplement, $n^{1/3}(\widehat{\xi}_n - \xi_0)$ converges in distribution to $\arg max_t Z^*(t)$, where the process $Z^*(t) = -\frac{1}{2}t^T V^* t + W^*(t)$ V^* is an $l \times l$ positive definite matrix and $W^*(t)$ is a centered Gaussian process with continuous sample paths and covariance kernel function $K^*(C_1, C_2)$. The expressions for V^* and $K^*(\cdot, \cdot)$ are given in the online supplement.

5 Numerical results

5.1 Simulations

Example 1 (single-stage optimal treatment regime)—We compare estimating the conventional mean-optimal treatment regime and quantile-optimal treatment regime in this example. We generate random data from the model

$$Y = 1 + X_1 - X_2 + X_3^3 + e^{X_4} + A\left(3 - 5X_1 + 2X_2 - 3X_3 + X_4\right) + (1 + A(1 + X_1 + X_2 + X_3 + X_4))\varepsilon,$$

where X_k ($k = 1,\dots, 4$) are independent Uniform(0, 1) random variables and $\varepsilon \sim N(0, 1)$ is independent of the covariates. The binary treatment indicator A satisfies log ($P(A = 1|X)/P(A_i = 0|X)) = -0.5 - 0.5\ (X_1 + X_2 + X_3 + X_4)$, where $X = (X_1,\dots, X_4)'$.

We consider the class of treatment regimes $I(\eta_0 + \eta^T X > 0)$, where $(\eta_0, \eta_1,\dots, \eta_4)^T$ has L_2-norm 1. Let $\mu(a, X) = E(Y/A = a, X)$, where $a \in \{0, 1\}$. The mean optimal treatment regime is given by $I(\mu(1, X) > \mu(0, X))$. In our example, it is $I(3 - 5X_1 + 2X_2 - 3X_3 + X_4 > 0)$, which belongs to our class of candidate treatment regimes. We compare the proposed method with two popular methods for estimating the mean-optimal treatment regime: a model-based approach and a model-free approach. For the model-based approach we impose models for $\mu(a, X)$ and then estimate the mean-optimal treatment regime by $I(\hat{\mu}(1, X) > \hat{\mu}(0, X))$, where $\hat{\mu}$ is the estimated value of μ. We consider two posited models for $\mu(a, X)$: (1) correctly specified regression function $\mu_t(a, X) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3^3 + \gamma_4 e^{X_4} + a(\gamma_5 + \gamma_6 X_1 + \gamma_7 X_2 + \gamma_8 X_3 + \gamma_9 X_4)$; and (2) misspecified regression function $\mu_m(a, X) = \exp[\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3^3 + a(\gamma_4 + \gamma_5 X_1 + \gamma_6 X_2 + \gamma_7 X_3 + \gamma_8 X_4)]$. For the model-free approach, we consider the estimator in Zhang et al. (2012a). We denote these three estimators by $\text{mean\_}RG_{\mu_t}$, $\text{mean\_}RG_{\mu_m}$ and mean_ZTLD, respectively.

For the quantile criteria, we consider $\tau = 0.25$ and 0.1, and denote the corresponding criterion as 0.25qt criterion and 0.10qt criterion, respectively. We do not have a closed form expression for the quantile-optimal treatment regime. In Table 2, based on a Monte Carlo experiment with sample size 10^5, we report the values of the ηi's indexing the optimal treatment regimes corresponding to different criteria; the last three columns of the table contain the mean, the 0.25 quantile and the 0.1 quantile of the outcomes if the corresponding optimal treatment regime is applied. These values will serve as our gold standard.

Table 3 summarizes the estimated optimal treatment regimes corresponding to the mean criterion (using $\text{mean\_}RG_{\mu_t}$, $\text{mean\_}RG_{\mu_m}$ and mean_ZTLD, respectively), the 0.25qt criterion and the 0.10qt criterion for sample sizes n =500 and 1000. The last three columns of Table 3 report the estimated mean, the 0.25 quantile and the 0.1 quantile of the outcomes if the estimated optimal treatment regime is applied. We observe the model-based approach for estimating the mean-optimal treatment regime is sensitive to the specified regression model and can be biased when the regression model is misspecified ($\text{mean\_}RG_{\mu_m}$ gives very biased estimators for η_0 and η_4). Also, the estimated optimal treatment regimes (and their achievable performance in terms of the value of the criterion functions) using the model-free approach get closer to the ideal ones reported in Table 2 as the sample size increases.

Example 2 (two-stage DTR)—We generate random data from the following model $Y = 1 + X_1 + A_1 [1 - 3 (X_1 - 0.2)^2] + X_2 + A_2 [1 - 5 (X_2 - 0.4)^2] + (1 + 0.5A_1 - A_1X_1 + 0.5A_2 - A_2X_2)\varepsilon$, where $\varepsilon \sim N(0, 0.4)$ $X_1 \sim$ Uniform $(0, 1)$ $X_2|\{X_1, A_1\} \sim$ Uniform$(0.5X_1, 0.5X_1 + 0.5)$, $A_1|X_1 \sim$ Bernoulli (expit $(-0.5 + X_1)$), and $A_2|\{X_1, A_1, X_2\} \sim$ Bernoulli (expit $(-1 + X_2)$) with expit $(t) = e^t/(1 + e^t)$. We consider sequential treatment regimes of the form (A_1, A_2), where $A_1 = I\{X_1 < \eta_1\}$ and $A_2 = I\{X_2 < \eta_2\}$. We note that this class contains the mean-optimal sequential treatment regimes which are given by $A_1 = I(X_1 < 0.777)$ and $A_2 = I(X_2 < 0.847)$.

The popular Q-learning procedure relies on specification of models for the so-called Q-functions. In this example, we compare with standard application of Q-learning based on linear models, that is, the Q-functions are specifies as $Q_t(H_t, A_t, \beta_t) = H_{t,0}^T\beta_{t,0} + A_tH_{t,1}^T\beta_{t,1}, t = 1, 2$, where $H_{2,0} = (1, X_1, A_1, X_1A_1, X_2)^T$, $H_{2,1} = (1, X_2)^T$, $H_{1,0} = (1, X_1)^T$; and $H_{1,1} = (1, X_1)^T$. We note that in practice the Q-learning procedure usually misspecifies the Q-function. We also compare with the model-free approach for estimating the mean-optimal dynamic treatment regime (Zhang et al. (2013)).

Table 4 reports the parameters indexing the optimal treatment regimes and the corresponding mean, median and 0.75 quantile of the outcome if the optimal treatment regime is applied, based on a Monte Carlo experiment with sample size 10^5. Table 5 summarizes the estimated parameters indexing the optimal treatment regimes and their estimated performance corresponding to different criteria for sample sizes $n = 500$, 1000, based on 400 simulation runs. The estimated optimal treatment regimes and their achievable performance are quite close to the ideal ones reported in Table 4, particularly when the sample size is large.

5.2 ACTG175 data analysis

We illustrate the proposed quantile-optimal treatment regime estimation method on the ACTG175 data set from the R package `speff2trial`, which contains measurements on 2139 HIV-infected patients. The patients were randomized to four treatment arms: zidovudine (AZT) monotherapy, AZT+didanosine (ddI), AZT+zalcitabine(ddC), and ddI monotherapy. The goal of the original clinical trial was to evaluate whether treatment of HIV infection with one drug (monotherapy) was the same as, better than, or worse than treatment with two drugs (combination therapy) in patients with CD4 T cells between 200 and $500/mm^3$ (Hammer et al., 1996). Figures 1 and 2 of the online supplement display the histograms of the response variable (CD4 count at week 96) for each of the two treatment arms for different subgroups of patients for which the subgroups are formed by the observed values of the CD4 count at week 0 or baseline weight. The varying shapes of the histograms across different ranges of both covariates indicate heteroscedastic treatment effects. It is also observed that the distribution of the response variable tends to be asymmetric and skewed to the right.

A basic conclusion from the study is for patients who had taken AZT before entering the trial, treatments with ddI or AZT – ddI are better than continuing to take AZT alone. There are $n = 562$ patients with full CD4 information that had taken AZT before the study and received AZT+ddI or ddI monotheraphy in this trial. Motivated by the aforementioned finding, we consider the problem of how to assign treatment to the patients who had taken AZT before, either to the AZT+ddI combination therapy or to the ddI monotheraphy. The quantitative outcome is the CD4 count at 96± 5 weeks from baseline (denoted as cd496) as CD4 count represents a vital signal for disease progression for HIV-infected patients. The treatment indicator A_i is set to 1 if patient i is assigned to the AZT + ddI therapy, and A_i is set to 0 if the patient is assigned to the ddI monotheraphy. Because this trial is randomized, the propensity score $\pi_i = n^{-1}\, \Sigma A_i = 0.48$ is taken as a constant for all subjects.

Two covariates are considered for estimating the optimal treatment regimes: X_1 (baseline weight of patient, measured in kg) and X_2 (baseline CD4 T cell count, denoted by cd40). It has been observed that body weight has a significant role on AZT pharmacokinetic profile. Burger et al. (1994) reported that AZT clearance is significantly lower in patients with a lower body weight, which indicates a qualitative interaction with AZT. In medicine, drug clearance is a pharmacokinetic measurement of the rate at which the active drug is removed from the body, and drug clearance is correlated with the time course of a drug's action (Hammer et al., 1996).

Let $X = (X_1, X_2)$, where both X_1 and X_2 are standardized to the interval [0, 1]. We consider the class of candidate regimes of the form $I\{\eta_0 + \eta_1 X_1 + \eta_2 X_2 < 0\}$, where $(\eta_0, \eta_1, \eta_2) \in (-1, 1)^3$. When the decision rule takes the value 1, the patient is assigned to the AZT+ddI combination therapy; otherwise the patient is assigned to the ddI monotheraphy. For identifiabilitv, we impose the restriction $\| \eta \| = 1$. We estimate the optimal treatment regimes using the median criterion, quartile criterion and the mean criterion. The median criterion is motivated by the robustness consideration; the quartile criterion is motivated by the desire to improve the treatment effect for weaker patients. Table 6 summaries the estimated optimal treatment regimes for the three criteria.

The estimated median of the potential outcome when the median-optimal treatment regime is applied is 360; whereas the median of the observed outcome is 339.5. The estimated first quartile of the potential outcome when the 0.25qt criterion is applied is 263; whereas the 0.25 quartile of the observed outcome is 237. The estimated mean of the potential outcome when the mean-optimal treatment regime is applied is 403.9; whereas the mean of the observed outcome is 355. Figure 3 of the online supplement depicts the three estimated regimes graphically, from which we observe that they are dramatically different from each other.

6 Conclusions and discussions

In a variety of applications, it is of interest to consider a treatment regime that maximizes the median or other quantile of the potential outcome distribution. This paper studies robust estimation of quantile-optimal static/dynamic treatment regimes. We propose a novel representation that expresses the parameter indexing the optimal treatment regime as a solution to an optimization problem with a nuisance parameter. Employing this representation and empirical process techniques, we prove that the estimated parameter indexing the quantile-optimal treatment regime has a nonstandard convergence rate and a non-normal limiting distribution. Our approach does not rely on the specification of an outcome regression model. We also investigate the doubly robust estimator for the quantile-optimal treatment regime, which can improve the estimation efficiency when a reliable outcome regression model is available (Section 1.1 of the online supplement).

Our proposed novel representation applies to a general class of policy search estimators for optimal treatment regimes defined by a general class of criteria. In particular, our approach can be applied to investigate the asymptotic distribution for the estimators of the mean-optimal treatment regimes in Zhang et al. (2012a, 2013) and fill in an important gap in the theory. The aforementioned nonstandard asymptotics will also arise when the mean-optimal criterion is used. For alternative criteria, we discuss optimal treatment regimes defined by the Ginhs mean difference criterion and the weighted quantile criterion in the online supplement, where an outline of the theory and some numerical examples are provided.

It is worth noting that the nonstandard asymptotics discussed in this paper are different from the nonregular asymptotics for Q-learning estimators. The Q-learning method models the stage-specific conditional mean functions and is a popular indirect method for estimating mean-optimal treatment regimes. Consider the Q-learning method in a two-stage dynamic setting and denote the estimated parameters indexing the optimal treatment regimes for the two stages as $(\hat{\psi}_1, \hat{\psi}_2)$. The asymptotic distribution for $\hat{\psi}_2$ is standard but the asymptotic distribution for $\hat{\psi}_1$ is nonregular in the sense that it does not converge uniformly over the parameter space (Robins, 2004; Chakraborty et al., 2010; Laber et al., 2014). The asymptotic distribution of $\hat{\psi}_1$ can change abruptly from being asymptotically normal to being non-normal depending on whether a certain event occurs with probability zero or not. This happens because $\hat{\psi}_1$ is a nonsmooth function of $\hat{\psi}_2$. The results in this paper and those in the literature on Q-learning demonstrate the challenges of asymptotic theory for optimal treatment regimes estimation. In general, classical asymptotic theory is no longer applicable.

An interesting future research direction is to investigate estimating quantile-optimal treatment regimes for survival data, where the response variable is randomly censored. Censored data arise in diverse fields such as economics, medicine and sociology. For example, in a clinical trial censoring occurs when a study ends before all patients experience the event of interest. Several authors (Goldberg and Kosorok (2012); Zhao et al. (2015c); Geng et al. (2015); Jiang et al. (2016)) recently studied estimating optimal treatment regimes with survival outcomes but have not considered the quantile criterion. When censoring is heavy, it can be difficult to estimate the mean survival time accurately but it is often possible to reliably estimate the median and the lower quantiles.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Sherwood's work was done when he was a graduate student at University of Minnesota. Wang's research is partly supported by NSF DMS-1512267 and DMS-1712706. Song's research is partly supported by NSF DMS-1555244 and NCI P01 CA142538. We thank the Co-Editor Nicholas Jewell, the AE and two anonymous referees for their constructive comments which help us significantly improve the paper. We also thank Dr. Shannon Holloway at North Carolina State University for proofreading the paper and providing many helpful comments.

Appendix: Technical Proofs

We provide below the proofs of Lemma 2 and Theorem 1. The proofs of Lemma 1, Theorem 2, and derivation of the theory for Section 6.2 are given in the online supplement.

Proof of Lemma 2

(1) Note that $g(\cdot, \beta, m) = [AI(X^T\beta > 0) + (1 - A)I(X^T\beta \le 0)]I(Y - m > 0)$. The classes $\left\{I(X^T\beta > 0) : \beta \in \mathbb{B}\right\}$ and $\{I(Y - m > 0) : m \in \mathbb{R}\}$ are both VC subgraph classes and hence bounded Donsker classes. Therefore, the class $\{g(\cdot, \beta, m) : \beta \in \mathbb{B}, m \in \mathbb{R}\}$ is Donsker (van der Vaart and Wellner (1996)). We thus have

$$\sup_{\beta \in \mathbb{B}, m \in \mathbb{R}} |P_n g(\cdot, \beta, m) - Pg(\cdot, \beta, m)| = O_p(n^{-1/2}). \quad (8)$$

We denote the supremum at the left side of the above expression as Δ_n. For any given β, $Pg(\cdot, \beta, m)$ is a decreasing function of m. Hence the assumption about the density ensures that there exists a constant $\kappa_1 > 0$ such that $\sup_{\beta \in \mathbb{B}} P \quad g(\cdot, \beta, m_0 + \varepsilon) < \frac{1 - \tau}{2} - \kappa_1 \varepsilon$, for each small enough $\varepsilon > 0$. Taking $\varepsilon = \Delta_n / \kappa_1$, to all n sufficiently large, it follows from (8) that $\sup_{\beta \in \mathbb{B}} P_n g(\cdot, \beta, m_0 + \Delta_n / \kappa_1) < \Delta_n + \frac{1 - \tau}{2} - \kappa_1 \frac{\Delta_n}{\kappa_1} = \frac{1 - \tau}{2}$. This implies $\hat{m} < m_0 + \Delta_n / \kappa_1$ for all n sufficiently large. Similarly, there exists a constant $\kappa_2 > 0$ such that $\sup_{\beta \in \mathbb{B}} Pg(\cdot, \beta, m_0 - \varepsilon) \ge \frac{1 - \tau}{2} + \kappa_2 \varepsilon$, for all small enough $\varepsilon > 0$. If follows that $\sup_{\beta \in \mathbb{B}} P_n g(\cdot, \beta, m_0 - \Delta_n / \kappa_2) \ge -\Delta_n + \frac{1 - \tau}{2} - \kappa_2 \frac{\Delta_n}{\kappa_2} = (1 - \tau)/2$ for all n sufficiently large. This implies $\hat{m}_n \ge m_0 - \Delta_n / \kappa_2$ for all n sufficiently large. Since $\Delta_n = O_p(n^{-1/2})$, we have $\hat{m}_n = m_0 + O_p(n^{-1/2})$. (2) Observing (i) $\hat{\beta}_n = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}} P_n g(\cdot, \beta, \hat{m}_n)$, (ii) $\beta = \beta_0$ uniquely maximizes $Pg(\cdot, \beta, m_0)$ and (iii) $\sup_{\beta \in \mathbb{B}} |P_n g(\cdot, \beta, \hat{m}_n) - Pg(\cdot, \beta, m_0)| = o_p(1)$, we conclude that $\hat{\beta}$ is consistent for β_0 by applying standard arguments of the M estimation theory (simple modification of Theorem 5.7 in van der Vaart (1998)). Next, we will show $\hat{\beta}_n - \beta_0 = O_p(n^{-1/3})$.

Let $\theta = (\beta^T, \delta)^T$, where $\delta = m - m_0$, and $h(\cdot, \beta, \delta) = C(\beta)I\{Y - m_0 - \delta > 0\} - C(\beta_0)I\{Y - m_0 - \delta > 0\}$. By definition, $\hat{\beta}_n = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}} P_n h(\cdot, \beta, \hat{m}_n - m_0)$. We will consider a Taylor expansion of $Ph(\cdot, \beta, \delta)$ around $\theta_0 = (\beta_0^T, 0)^T$. Note that $h(\cdot, \beta_0, 0) = 0$ and that

$$E[C(\beta)I\left\{Y - m_0 - \delta > 0\right\}] = \frac{1}{2}E\left\{I(X^T\beta > 0)I(Y - m_0 - \delta > 0)|A = 1\right\} + \frac{1}{2}E\left\{I(X^T\beta \le 0)I(Y - m_0 - \delta > 0)|A = 0\right\} = \frac{1}{2}E\left\{I(X^T\beta > 0)S_{1,X}(m_0 + \delta)\right\} + \frac{1}{2}E\left\{I(X^T\beta \le 0)S_{0,X}(m_0 + \delta)\right\} = \frac{1}{2}E\left\{I(X^T\beta > 0)(S_{1,X}(m_0 + \delta) - S_{0,X}(m_0 + \delta))\right\} + \frac{1}{2}E\left\{S_{0,X}(m_0 + \delta)\right\},$$

where $S_{1,X}(\cdot)$ and $S_{0,X}(\cdot)$ are the conditional survival functions of $Y^*(1)$ and $Y^*(0)$ given X, respectively. Let $q(X, \delta) = S_{1,X}(m_0 + \delta) - S_{0,X}(m_0 + \delta)$, then

$$E(h(\cdot, \beta, \delta)) = \frac{1}{2}E\left\{\left(I(X^T\beta > 0) - I(X^T\beta_0 > 0)\right)q(X, \delta)\right\}.$$

It is easy to see $\frac{\partial}{\partial\delta}E(h(\cdot, \beta, \delta))|_{\beta = \beta_0, \delta = 0} = 0$ and $\frac{\partial^2}{\partial\delta^2}E(h(\cdot, \beta, \delta))|_{\beta = \beta_0, \delta = 0} = 0$. Note that the transformation $T_\beta = (I - ||\beta||^{-2}\beta\beta^T)(I - \beta_0\beta_0^T) + ||\beta||^{-1}\beta\beta_0^T$, where I denotes the identity matrix, maps the region $A = \{X^T\beta_0 > 0\}$ onto $A(\beta) = \{X^T\beta > 0\}$, taking ∂A to $\partial A(\beta)$. The surface measure σ_β on $\partial A(\beta)$ has the constant density $\rho_\beta(X) = \beta^T\beta_0/||\beta||$ with respect to the image of the surface measure $\sigma = \sigma_{\beta 0}$ under T_β. The outward pointing unit vector normal to $A(\beta)$ is the standardized vector $-\beta/||\beta||$ and along ∂A the derivative $(\partial/\partial\beta)T_\beta(X)$ reduces to $-||\beta||^{-2}[\beta X^T + (\beta^T X)I]$. Using the result from Section 10.7 of Loomis and Sternberg (1968) on derivatives as surface integrals, we have

$$\frac{\partial}{\partial\beta^T}E(h(\cdot, \beta, \delta)) = \frac{1}{2}||\beta||^{-2}\beta^T\beta_0(I + ||\beta||^{-2}\beta\beta^T)\int I\left\{X^T\beta_0 = 0\right\}q(T_\beta(X), \delta)f(T_\beta(X))Xd\sigma.$$

Note that we have $\frac{\partial}{\partial\beta}E(h(\cdot, \beta, \delta))|_{\beta = \beta_0, \delta = 0} = 0$because $E(h(\cdot, \beta, 0))$ is maximized at $\beta = \beta_0$. Combining with the observation that $T_{\beta 0}(X) = X$ along $\{X^T\beta_0 = 0\}$, we have $\int I\{X^T\beta_0 = 0\}I(X, 0)f(X)X\,d\sigma = 0$ Using this and the fact $||\beta_0|| = 1$, we have

$$\frac{\partial^2}{\partial\beta\partial\beta^T}E(h(\cdot, \beta, \delta))|_{\beta = \beta_0, \delta = 0} = -\frac{1}{2}\int I\left\{X^T\beta_0 = 0\right\}(f(X)\dot{q}(X, 0) + q(X, 0)\dot{f}(X))^T\beta_0 XX^T d\sigma,$$

where $\dot{q}(X, 0)$ and $\dot{f}(X)$ denote the gradients with respect to X. Also,

$$\frac{\partial^2}{\partial\beta^T\partial\delta}E(h(\cdot, \beta, \delta))|_{\beta = \beta_0, \delta = 0} = \frac{1}{2}\int I\left\{X^T\beta_0 = 0\right\}(s_{1,X}(m_0) - s_{0,X}(m_0))f(X)Xd\sigma,$$

where $s_{1,X}$ and $s_{0,X}$ are the derivatives of $S_{1,X}$ and $S_{0,X}$ with respect to δ, respectively. We write

$$V = -\frac{\partial^2}{\partial\beta\partial\beta^T}E(h(\cdot, \beta, \delta))|_{\beta = \beta_0, \delta = 0} \quad (9)$$

and $a_1 = \frac{\partial^2}{\partial \beta^T \partial \delta} E(h(\cdot, \beta, \delta))|_{\beta = \beta_0, \delta = 0}$, then the Taylor expansion of $Ph(\cdot, \beta, \delta)$ around (β_0, 0) has the form

$$Ph(\cdot, \beta, \delta) = -\frac{1}{2}(\beta - \beta_0)^T V(\beta - \beta_0) + a_1^T(\beta - \beta_0)\delta + o(||\beta - \beta_0||^2) + o(\delta^2). \quad (10)$$

For a given positive constant R, let $H_R = \sup_{||\theta - \theta_0|| \le R}|h(\cdot, \beta, \delta)|$. We observe that $h(\cdot, \beta, \delta)$ is nonzero if and only if $C(\beta)$ and $C(\beta_0)$ take different values. Hence, $H_R \le \sup_{||\theta - \theta_0|| \le R}\left\{I(X^T\beta > 0 \ge X^T\beta_0) + I(X^T\beta_0 > 0 \ge X^T\beta)\right\}$. The envelope function H_R is bounded by an indicator function of a pair of multidimensional wedge shaped regions, each subtending an angle of order $O(R)$, from which we deduce that $E(H_R^2) = O(R)$. The conditions of Lemma 4.1 of Kim and Pollard (1990) are satisfied. Hence, for each fixed $\varepsilon > 0$, uniformly for $\|\theta - \theta_0\| \le R$, $P_nh(\cdot, \beta, \delta) \le Ph(\cdot, \beta, \delta) + \varepsilon(\|\beta - \beta_0\|^2 + \delta^2) + O_p(n^{-2/3})$. Combining with the upper bound in (10), we have $P_nh(\cdot, \beta, \delta) \le -\left(\frac{1}{2}\lambda_{min}(V) - \varepsilon\right)||\beta - \beta_0||^2 + ||a_1||\,||\beta - \beta_0||\,|\delta| + (\varepsilon + o(1))\delta^2 + O_p\left(n^{-2/3}\right)$, where $\lambda_{min}(V)$ denotes the smallest eigenvalue of V. Choosing $\varepsilon = \lambda_{min}(V)/4$, we derive that

$$0 = P_nh\left(\cdot, \beta_0, \hat{m}_n - m_0\right) \le P_nh\left(\cdot, \hat{\beta}_n, \hat{m}_n - m_0\right) \le -\frac{1}{4}\lambda_{min}(V)||\hat{\beta}_n - \beta_0||^2 + O_p(n^{-1/2})||\hat{\beta}_n - \beta_0|| + O_p(n^{-2/3}).$$

Completing the square in $||\hat{\beta}_n - \beta_0||$, we derive that $||\hat{\beta}_n - \beta_0|| = O_p(n^{-1/3})$.

Next, we show that $\hat{\beta}_n$ nearly maximizes $P_nh(\cdot, \beta, 0)$. A similar argument as above shows that $P|h(\cdot, \theta_1) - h(\cdot, \theta_2)| = O(||\theta_1 - \theta_2||)$ for θ_1, θ_2 near θ_0. It follows from Lemma 4.6 of Kim and Pollard (1990) that the process $J_n(\cdot, \alpha, \gamma) = n^{2/3}(P_n - P)h(\cdot, \beta_0 + \alpha n^{-1/3}, \gamma n^{-1/3})$ satisfies the stochastic equicontinuity condition of Theorem 2.3 of Kim and Pollard (1990). Since $n^{1/3}(\hat{m}_n - m_0) = o_p(1)$, this implies that for β uniformly in a $O(n^{-1/3})$ neighborhood of β_0, $J_n(\cdot, n^{1/3}(\beta - \beta_0), n^{1/3}(\hat{m}_n - m_0)) - J_n(\cdot, n^{1/3}(\beta - \beta_0), 0) = o_p(1)$. That is, $P_nh(\cdot, \beta, \hat{m}_n - m_0) = P_nh(\cdot, \beta, 0) + Ph(\cdot, \beta, \hat{m}_n - m_0) - Ph(\cdot, \beta, 0) + o_p(n^{-2/3})$, uniformly over an $O_p(n^{-1/3})$ neighborhood of β_0. Within such a neighborhood, Taylor expansion similarly as before shows that $Ph(\cdot, \beta, \hat{m}_n - m_0) - Ph(\cdot, \beta, 0) = o_p(n^{-2/3})$. Suppose $\tilde{\beta}_n = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}}\, P_nh(\cdot, \beta, 0)$. An analysis similar to that for $\hat{\beta}_n$ shows that $\tilde{\beta}_n = O_p(n^{-1/3})$.

Hence,

$$P_n h(\cdot, \hat{\beta}_n, 0) = P_n h(\cdot, \hat{\beta}_n, \hat{m}_n - m_0) - o_p(n^{-2/3}) \geq P_n h(\cdot, \tilde{\beta}_n, \hat{m}_n - m_0) - o_p(n^{-2/3}) = P_n h(\cdot, \tilde{\beta}_n, 0) - o_p(n^{-2/3}),$$

where the inequality follows because $\hat{\beta}_n = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}}\, P_n h(\cdot, \beta, \hat{m}_n - m_0)$. Therefore, $P_n h(\cdot, \hat{\beta}_n 0) \geq \sup_{\beta \in \mathbb{B}} P_n h(\cdot, \beta, 0) - o_p(n^{-2/3})$. □

Proof of Theorem 1

Following Lemma 2(2), to find the asymptotic distribution of $n^{1/3}(\hat{\beta}_n - \beta)$, it suffices to apply the main theorem of Kim and Pollard (1990) to the one parameter process $\{h(\cdot, \beta, 0): \beta \in \mathbb{B}\}$. Recall that $h(\cdot, \beta, 0) = C(\beta)I\{Y > \mathrm{m}_0\} - C(\beta_0)I\{Y > m_0\}$. In the following, we will verify conditions (iv) and (v) of the main theorem of Kim and Pollard (1990). Other conditions of the theorem are relatively easier and can be checked using similar techniques as those in the proof of Lemma 2.

For condition (iv), it can be shown that $\frac{\partial^2}{\partial\beta\partial\beta^T} E(h(\cdot, \beta, 0))|_{\beta = \beta_0} = -V$, where V is defined in (9) in the proof of Lemma 2. Next, we calculate the kernel function in condition (v). Similarly as in the calculation in the proof of Lemma 2, for each C_1, C_2 in R^l, and $t > 0$,

$$\begin{aligned}
& tP|h\left(\cdot, \beta_0 + \frac{C_1}{t}, 0\right) - h\left(\cdot, \beta_0 + \frac{C_2}{t}, 0\right)|^2 \\
&= tP\left\{|C(\beta_0 + C_1/t) - C(\beta_0 + C_2/t)|I(Y > m_0)\right\} \\
&= \frac{1}{2} tP\left\{|I(X^T(\beta_0 + C_1/t) > 0) - I(X^T(\beta_0 + C_2/t) > 0)|I(Y^*(1) > m_0)\right\} + \frac{1}{2} tP\left\{|I(X^T(\beta_0 + C_1/t) \leq 0) - I(X^T(\beta_0 + C_2/t) \leq 0)|I(Y^*(0) > m_0)\right\} \\
&= tP\left\{(S_{1,X}(m_0) + S_{0,X}(m_0))|I(X^T(\beta_0 + C_1/t) > 0 - I(X^T(\beta_0 + C_2/t) > 0)|\right\}.
\end{aligned}$$

To evaluate the above expression, we make use of the local coordinates (Example 6.4 of Kim and Pollard (1990)), for which we define $\beta(\tau) = \sqrt{1 - ||\tau||^2}\beta_0 + \tau$, where $\boldsymbol{\tau}$ is orthogonal to β_0 and ranges over a neighborhood of the origin. It is noted that as the parameter space is on the sphere ($||\beta_0|| = 1$, $||\beta|| = 1$), such a decomposition can be obtained by taking $\boldsymbol{\tau} = \tau(\beta) = T_0\beta$, where $T_0 = I - \beta_0\beta_0^T$ Then we can write $\beta = (\beta_0^T\beta)\beta_0 + T_0\beta$ such that $\beta_0^T\beta = \sqrt{1 - ||\tau||^2}$ and $\beta_0^T T_0 \beta = 0$. Also, $\tau(\beta_0 + C_1/t) = T_0 C_1/t$, $\tau(\beta_0 + C_2/t) = T_0 C_2/t$. Similarly, we can decompose X as $X = r\beta_0 + Z$ for some random variable r and random vector Z, with Z being orthogonal to β_0. Let $C_k^* = T_0 C_k \in T_0$, $k = 1,2$, then $X^T(\beta_0 + C_1/t) = (r\beta_0 + Z)^T(\sqrt{1 - ||C_1^*||^2}\beta_0 + C_1^*/t) = r\sqrt{1 - ||C_1^*||^2/t^2} + Z^T C_1^*/t$. Let $p(\cdot,\cdot)$ be the joint density function of (r, Z), which can be deduced from the density of X, With a change of variable $w = tr$, $tP\{(S_{1,X}(m_0) + S_{0,X}(m_0))|I(X^T(\beta_0 + C_1/t) > 0) - I(X^T(\beta_0 + C_2/t) > 0)|\}$ is equal to

$$\iint I\left\{-Z^T C_2^*(1-||C_2^*||^2/t^2)^{-1/2} > w \geq -Z^T C_1^*(1-||C_1^*||^2/t^2)^{-1/2}\right\}(S_{1,\frac{w}{t}\beta_0+Z}(m_0) + S_{0,\frac{w}{t}\beta_0+Z}(m_0))p(w/t,Z)dwdZ + \iint I\left\{-Z^T C_1^*(1-||C_1^*||^2/t^2)^{-1/2} > w \geq -Z^T C_2^*(1-||C_2^*||^2/t^2)^{-1/2}\right\}\left(S_{1,\frac{w}{t}\beta_0+Z}(m_0) + S_{0,\frac{w}{t}\beta_0+Z}(m_0)\right)p(w/t,Z)dwdZ.$$

Integrating over w and letting $t \rightarrow \infty$to get

$\lim_{t \to \infty} tP|h(\cdot,\beta_0+C_1/t,0)-h(\cdot,\beta_0+C_2/t,0)|^2$. We denote this limit as $L(C_1-C_2)$. Using

$$= \int |Z^T(C_1^*-C_2^*|)|(S_{1,Z}(m_0)+S_{0,Z}(m_0))p(0,Z)dZ$$

$$= \int |Z^T(C_1-C_2)|(S_{1,Z}(m_0))+S_{0,Z}p(m_0))p(0,Z)dZ$$

the identity $2xy = x^2+y^2-(x-y)^2$, we deduce that the limiting covariance kernel function can be written as

$$K(C_1,C_2) = \lim_{t \to \infty} tP\left\{h(\cdot,\beta_0+C_1/t,0)-h(\cdot,\beta_0+C_2/t/t,0) = \lim_{t \to \infty}\frac{1}{2}tP|h(\cdot,\beta_0+C_1/t,0).\right. - h(\cdot,\beta_0,0)|^2 + \lim_{t \to \infty}\frac{1}{2}tP|h(\cdot,\beta_0+C_2/t,0)-h(\cdot,\beta_0,0)|^2 - \lim_{t \to \infty}\frac{1}{2}\lim_{t \to \infty}tP|h(\cdot,\beta_0 + C_1/t,0)-h(\cdot,\beta_0+C_2/t,0)|^2 = \frac{1}{2}(L(C_1)+L(C_2)-L(C_1-C_2))$$

The asymptotic distribution of $n^{1/3}(\hat{\beta}_n-\beta_0)$ then follows by applying the main theorem of Kim and Pollard (1990) □

References

Behncke S, Froelich M, Lechner M. Targeting labour market programmes: Results from a randomized experiment. Swiss Journal of Economics and Statistics. 2009; 145(3):221–268.

Bhattacharya D. Inferring optimal peer assignment from experimental data. Journal of the American Statistical Association. 2009; 104(486):486–500.

Bhattacharya D, Dupas P. Inferring welfare maximizing treatment assignment under budget constraints. Journal of Econometrics. 2012; 167(1):168–196.

Burger DM, Meenhorst PL, ten Napel CH, Mulder JW, Neef C, Koks CH, Bult A, Beijnen JH. Pharmacokinetic variability of zidovudine in hiv-infected individuals: subgroup analysis and drug interactions. AIDS. 1994; 8(12):1683–1690. [PubMed: 7888117]

Cai T, Tian L, Wong PH, Wei L. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011; 12(2):270–282. [PubMed: 20876663]

Chakraborty B, Moodie EE. Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine. Springer Science & Business Media; 2013.

Chakraborty B, Murphy S, Strecher V. Inference for non-regular parameters in optimal dynamic treatment regimes. Statistical Methods in Medical Research. 2010; 19(3):317–343. [PubMed: 19608604]

Chakraborty B, Murphy SA. Dynamic treatment regimes. Annual Review of Statistics and its Application. 2014; 1:447.

Chernozhukov V, Hansen C. An iv model of quantile treatment effects. Econometrica. 2005; 73(1): 245–261.

Dehejia RH. Program evaluation as a decision problem. Journal of Econometrics. 2005; 125(1):141–173.

Frölich M. Statistical treatment choice: an application to active labor market programs. Journal of the American Statistical Association. 2008; 103:547–558.

Geng Y, Zhang HH, Lu W. On optimal treatment regimes selection for mean survival time. Statistics in medicine. 2015; 34(7):1169–1184. [PubMed: 25515005]

Gerber AS, Green DP. The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. American Political Science. 2000; 94:653–663. Review.

Goldberg Y, Kosorok MR. Q-learning with censored data. Annals of Statistics. 2012; 40(1):529. [PubMed: 22754029]

Hammer SM, Katzenstein DA, Hughes MD, Gundacker H, Schooley RT, Haubrich RH, Henry WK, Lederman MM, Phair JP, Niu M, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. New England Journal of Medicine. 1996; 335(15):1081–1090. [PubMed: 8813038]

Henderson R, Ansell P, Alshibani D. Regret-regression for optimal dynamic treatment regimes. Biometrics. 2010; 66(4):1192–1201. [PubMed: 20002404]

Hirano K, Porter JR. Asymptotics for statistical treatment rules. Econometrica. 2009; 77(5):1683–1701.

Hogan JW, Lee JY. Marginal structural quantile models for longitudinal observational studies with time-varying treatment. Statistica Sinica. 2004:927–944.

Huang X, Choi S, Wang L, Thall PF. Optimization of multi-stage dynamic treatment regimes utilizing accumulated data. Statistics in medicine. 2015; 34(26):3424–3443. [PubMed: 26095711]

Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics. 2013; 7:443–470.

Jiang R, Lu W, Song R, Davidian M. On estimation of optimal treatment regimes for maximizing t-year survival probability. Journal of the Royal Statistical Society: Series B. 2016 In Press.

Kim JK, Pollard D. Cube root asymptotics. The Annals of Statistics. 1990; 1:191–219.

Kosorok MR, Moodie EE. ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, ASA; Alexandria, VA: 2016. Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine.

Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. Dynamic treatment regimes: Technical challenges and applications. Electronic Journal of Statistics. 2014; 8(1):1225. [PubMed: 25356091]

Lavori PW, Dawson R. A design for testing clinical strategies: biased adaptive within-subject randomization. Journal of the Royal Statistical Society: Series A. 2000; 163:29–38.

Linn KA, Laber EB, Stefanski LA. Interactive q-learning for probabilities and quantiles. 2015 arXiv 1407.3414.

Loomis LH, Sternberg S. Advanced Calculus. Addison-Wesley Reading; Mass: 1968.

Manski CF. Statistical treatment rules for heterogeneous populations. Econometrica. 2004; 72(4): 1221–1246.

Matsouaka RA, Li J, Cai T. Evaluating marker-guided treatment selection strategies. Biometrics. 2014; 70(3):489–499. [PubMed: 24779731]

Moodie E, Dean N, Sun Y. Q-learning: Flexible learning about useful utilities. Statistics in Biosciences. 2014; 6:223–243.

Moodie EE, Platt RW, Kramer MS. Estimating response-maximized decision rules with applications to breastfeeding. Journal of the American Statistical Association. 2009; 104:155–165.

Moodie EE, Richardson TS. Estimating optimal dynamic regimes: Correcting bias under the null. Scandinavian Journal of Statistics. 2010; 37(1):126–146.

Moodie EE, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics. 2007; 63(2):447–455. [PubMed: 17688497]

Murphy SA. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society: Series B. 2003; 65(2):331–366.

Murphy SA. An experimental design for the development of adaptive treatment strategies. Statistics in Medicine. 2005a; 24(10):1455–1481. [PubMed: 15586395]

Murphy SA. A generalization error for q-learning. Journal of Machine Learning Research. 2005b; 6:1073–1097. [PubMed: 16763665]

Murphy SA. An experimental design for the development of adaptive treatment strategies. Statistics in Medicine. 2008; 24:1455–1481.

Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science. 1990; 5(4):465–472.

Orellana LRA, Robins J. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. The International Journal of Biostatistics. 2010; 6

Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Statist. 2011; 39(2):1180–1210.

Qian M, Nahum-Shani I, Murphy SA. Modern Clinical Trial Analysis. Springer; 2012. Dynamic treatment regimes; 127–148.

Robins J, Hernan M, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11:550–560. [PubMed: 10955408]

Robins JM. Latent variable modeling and applications to causality (Los Angeles, CA, 1994), volume 120 of Lecture Notes in Statist. Springer; New York: 1997. Causal inference from complex longitudinal data; 69–117.

Robins JM. Proceedings of the Second Seattle Symposium in Biostatistics. Springer; 2004. Optimal structural nested models for optimal sequential decisions; 189–326.

Robins JM, O L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. Statistics in Medicine. 2008; 27:4678–4721. [PubMed: 18646286]

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.

Rousseeuw PJ. Least median of squares regression. Journal of the American Statistical Association. 1984; 79:871–880.

Rubin D. Estimating causal effects of treatments in randomized and non-randomized studies. Journal of educational Psychology. 1974; 66:688–701.

Rubin DB. Bayesian inference for causal effects: the role of randomization. The Annals of Statistics. 1978; 6:34–58.

Rubin DB. Which ifs have causal answers. Journal of the American Statistical Association. 1986; 81:961–962.

Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Q-and a-learning methods for estimating optimal dynamic treatment regimes. Statistical Science. 2014; 29(4):640. [PubMed: 25620840]

Song R, Wang W, Zeng D, Kosorok M. Penalized q-learning for dynamic treatment regimens. Statistica Sinica. 2015; 25:901–920. [PubMed: 26257504]

Staghøj J, Svarer M, Rosholm M. Choosing the best training programme: Is there a case for statistical treatment rules? Oxford Bulletin of Economics and Statistics. 2010; 72:172–201.

Stoye J. Minimax regret treatment choice with finite samples. Journal of Econometrics. 2009; 151(1): 70–81.

Tao Y, Wang L. Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. Biometrics. 2017; 73(1):145–155. [PubMed: 27213913]

Tetenov A. Statistical treatment choice based on asymmetric minimax regret criteria. Journal of Econometrics. 2012; 166(1):157–165.

Thall PF, Sung HG, Estey EH. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. Journal of the American Statistical Association. 2011; 97:29–39.

van der Laan MJ, Petersen ML, Joffe MM. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. Journal of Biostatistics. 2005; 1

van der Vaart A. Asymptotic Statistics. Cambridge University Press; 1998.

van der Vaart AW, Wellner JA. Weak Convergence and Empirical Processes with Applications to Statistics. Springer-Verlag; New York: 1996.

Wallace MP, Moodie EE. Personalizing medicine: a review of adaptive treatment strategies. Pharmacoepidemiology and Drug Safety. 2014; 23(6):580–585. [PubMed: 24700536]

Watkins C, Dayan P. Q-learning. Maching Learning. 1992; 8:279–292.

Wunsch C. Optimal use of labor market policies: the role of job search assistance. Review of Economics and Statistics. 2013; 95(3):1030–1045.

Zhang B, Tsiatis A, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika. 2013; 100:681–694.

Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012a; 68(4):1010–1018. [PubMed: 22550953]

Zhang Z, Chen Z, Troendle JF, Zhang J. Causal inference on quantiles with an obstetric application. Biometrics. 2012b; 68:697–706. [PubMed: 22150612]

Zhao Y, Zeng D, Laber E, Song R, Yuan M, Kosorok M. Doubly robust learning for estimating individualized treatment with censored data. Biometrika. 2015a; 102:151–168. [PubMed: 25937641]

Zhao YQ, Zeng D, Laber EB, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. Journal of the American Statistical Association. 2015b; 110:583–598. [PubMed: 26236062]

Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR. Doubly robust learning for estimating individualized treatment with censored data. Biometrika. 2015c; 102(1):151–168. [PubMed: 25937641]

Zhao YQ, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association. 2012; 107(499):1106–1118. [PubMed: 23630406]

Zhou Y, Wang L, Sherwood B, Song R. Quantoptr: Algorithms for quantile- and mean-optimal treatment regimes. 2017. https://CRAN.R-project.org/package=quantoptr

Table 1

Mean, 0.25 quantile and 0.10 quantile of the potential outcomes corresponding to six different treatment regimes (based on a Monte Carlo experiment with 10^6 observations).

Regime	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
mean	1.50	**2.40**	2.37	2.00	1.78	2.00	1.74
$Q_{0.25}$	0.80	1.10	**1.14**	1.01	0.91	−0.02	0.59
$Q_{0.10}$	0.16	−0.03	0.20	**0.33**	0.26	−2.29	−0.81

Table 2

Population parameters and summary values for optimal treatment regimes under different criteria for Example 1 based on a Monte Carlo experiment with $n = 10^5$.

	η_0	η_1	η_2	η_3	η_4	Q_{mean}	$Q_{0.25}$	$Q_{0.1}$
mean criterion	0.43	−0.72	0.29	−0.43	0.14	3.99	2.28	0.55
0.25qt criterion	0.42	−0.60	0.41	−0.43	−0.34	3.79	2.46	1.18
0.1qt criterion	0.27	−0.68	0.38	−0.43	−0.37	3.44	2.36	1.55

Columns 2–6 are values of the η_i's of the optimal treatment regimes corresponding to different criteria. The last three columns are the mean. 0.25 quantile and 0.1 quantile of the potential outcomes if the optimal treatment regime is applied.

Table 3

Estimated optimal treatment regimes (mean criterion, 0.25 quantile criterion and 0.1 quantile criterion) and their corresponding value functions for Example 1.

Method	**n**	$\hat{\eta}_0$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\eta}_4$	$\hat{Q}_{\text{mean}}$	$\hat{Q}_{0.25}$	$\hat{Q}_{0.1}$
mean_RG_{μ_t}	500	0.42 (0.10)	−0.71 (0.07)	0.28 (0.13)	−0.41 (0.11)	0.14 (0.12)	3.99 (0.21)	2.29 (0.19)	0.56 (0.40)
	1000	0.43 (0.06)	−0.71 (0.05)	0.29 (0.09)	−0.43 (0.08)	0.14 (0.09)	3.99 (0.14)	2.28 (0.13)	0.52 (0.27)
mean_RG_{μ_m}	500	0.26 (0.11)	−0.71 (0.08)	0.30 (0.12)	−0.38 (0.12)	0.37 (0.12)	3.96 (0.21)	2.23 (0.19)	0.65 (0.38)
	1000	0.27 (0.08)	−0.71 (0.06)	0.31 (0.09)	−0.39 (0.08)	0.37 (0.09)	3.97 (0.15)	2.22 (0.13)	0.62 (0.27)
Mean_ZTLD	500	0.36 (0.2)	−0.63 (0.14)	0.31 (0.24)	−0.38 (0.2)	0.12 (0.27)	4.31 (0.21)	2.31 (0.21)	0.63 (0.53)
	1000	0.38 (0.15)	−0.67 (0.11)	0.29 (0.19)	−0.4 (0.15)	0.17 (0.19)	4.18 (0.13)	2.29 (0.16)	0.6 (0.47)
0.25qt criterion	500	0.38 (0.15)	−0.57 (0.14)	0.37 (0.19)	−0.37 (0.18)	−0.31 (0.2)	3.85 (0.26)	2.65 (0.16)	1.3 (0.39)
	1000	0.4 (0.12)	−0.59 (0.12)	0.35 (0.17)	−0.43 (0.12)	−0.28 (0.15)	3.81 (0.18)	2.57 (0.11)	1.31 (0.28)
0.10qt criterion	500	0.24 (0.23)	−0.56 (0.2)	0.3 (0.25)	−0.4 (0.22)	−0.33 (0.25)	3.5 (0.26)	2.45 (0.16)	1.75 (0.15)
	1000	0.27 (0.18)	−0.61 (0.14)	0.32 (0.22)	−0.44 (0.15)	−0.33 (0.19)	3.47 (0.18)	2.42 (0.11)	1.68 (0.11)

The numbers in the parenthesis are standard deviations. The last three columns are the estimated mean, 0.25 quantile and 0.1 quantile of the potential outcome if the estimated optimal treatment regime is applied. The three methods mean_RG_{μ_t}, mean_RG_{μ_m} and mean_ZTLD denote the mean-optimal treatment regime estimators using the model-based approach with correctly specified regression model, the model-based approach with incorrectly specified regression model and the approach of Zhang et al. (2012a), respectively.

Table 4

Population parameters and summary values for optimal treatment regimes under different criteria for Example 2 based on a Monte Carlo experiment with $n = 10^5$.

Method	η_1	η_2	Q_{mean}	$Q_{0.50}$	$Q_{0.75}$
Mean criterion	0.777	0.847	3.331	3.323	3.821
0.50qt criterion	0.753	0.808	3.327	3.327	3.827
0.75qt criterion	0.729	0.795	3.322	3.325	3.828

Columns 2–3 are values of the η_i's of the optimal treatment regimes corresponding to different criteria. The last three columns are the mean, median and 0.75 quantile of the potential outcomes if the optimal treatment regime is applied.

Table 5

Estimated optimal treatment regimes and their corresponding estimated value functions under different criteria for Example 2.

Method	***n***	$\boldsymbol{\eta}_1$	$\boldsymbol{\eta}_1$	$\widehat{Q}_{mean}$	$\widehat{Q}_{0.50}$	$\widehat{Q}_{0.75}$
mean_Qlearning	500	0.755(0.041)	1.176(0.294)	3.319(0.090)	3.309(0.102)	3.815(0.122)
	1000	0.752(0.027)	1.131(0.144)	3.321(0.065)	3.305(0.07)	3.819(0.079)
mean_ZTLD	500	0.773(0.073)	0.846(0.067)	3.370(0.095)	3.376(0.097)	3.862(0.118)
	1000	0.768(0.055)	0.852(0.059)	3.356(0.065)	3.354(0.068)	3.848(0.081)
0.50qt criterion	500	0.751(0.08)	0.815(0.079)	3.357(0.090)	3.391(0.102)	3.858(0.119)
	1000	0.750(0.062)	0.813(0.069)	3.343(0.063)	3.366(0.068)	3.849(0.081)
0.75qt criterion	500	0.734(0.108)	0.785(0.103)	3.328(0.095)	3.331(0.109)	3.892(0.123)
	1000	0.723(0.084)	0.795(0.095)	3.322(0.067)	3.326(0.075)	3.865(0.077)

The numbers in the parenthesis are standard deviations. The last three columns are the estimated mean, median and 0.75 quantile of the potential outcome if the estimated optimal treatment regime is applied. The mean_Qlearning method stands for the mean-optimal treatment regime estimator using the Q-learning approach. The mean_ZTLD method is the mean-optimal treatment regime estimator using Zhang et al. (2013).

Table 6

Estimated optimal treatment regimes and summary values for ACTG175 data analysis.

Method	$\hat{\eta}_0$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{Q}_{0.50}$	$\hat{Q}_{0.25}$	$\hat{Q}_{\text{mean}}$
0.50qt criterion	−0.571	0.691	0.444	360	220	375.4
0.25qt criterion	−0.210	0.958	−0.194	333	263	346.5
Mean criterion	−0.526	0.799	0.292	331	219	403.9